DOI: https://doi.org/10.52756/ijerr.2025.v47.002 *Int. J. Exp. Res. Rev.*, Vol. 47: 12-24 (2025)

Original Article | *Peer Reviewed* | *Open Access*

International Journal of Experimental Research and Review (IJERR)
© Copyright by International Academic Publishing House (IAPH)
ISSN: 2455-4855 (Online)
www.iaph.in
ISSN 2455-4855

An Empirical Analysis of Deep Learning Models for Temperature Prediction

Amrita Sarkar[1]*, Vandana Bhattacharjee[2] and Prachi Pandey[1]

[1]Department of CSE, Birla Institute of Technology, Mesra (Lalpur), Ranchi-834001, Jharkhand, India; [2]Department of CSE, Birla Institute of Technology, Mesra, Ranchi-835215, Jharkhand, India

E-mail/Orcid Id:

AS, amritasarkar@bitmesra.ac.in, https://orcid.org/0000-0001-7415-2822; *VB,* vbhattacharya@bitmesra.ac.in, https://orcid.org/0000-0002-0680-2691; *PP,* prachipandeyk85@gmail.com, https://orcid.org/0009-0007-8683-0935

Article History:

Received: 19th Sep., 2024
Accepted: 12th Dec., 2024
Published: 30th Apr., 2025

Keywords:
Deep learning, Environmental data, Temperature prediction, Time series forecasting

How to cite this Article:
Amrita Sarkar, Vandana Bhattacharjee and Prachi Pandey (2025). An Empirical Analysis of Deep Learning Models for Temperature Prediction. *International Journal of Experimental Research and Review*, *47*, 12-24.
DOI: https://doi.org/10.52756/ijerr.2025.v47.002

Abstract: Accurate temperature prediction is critical in diverse areas, such as agriculture, disaster management, and urban planning, where understanding climatic patterns is essential. This study explores the application of advanced deep-learning models for temperature forecasting, focusing on the model's ability to establish complex relationships and temporal dependencies within climatic data. This study evaluates the performance of various deep-learning models for temperature prediction using environmental data. Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM), and Bidirectional GRU (BiGRU) models were developed and compared. The models were trained on meteorological data from Ranchi, India, spanning 2014-2024. Performance was assessed using Root Mean Square Error (RMSE), loss function analysis, and statistical significance testing. Results indicate that bidirectional architectures (BiLSTM and BiGRU) consistently outperformed unidirectional models. BiLSTM achieved the lowest RMSE and most balanced loss values across training, validation and test sets. The BiLSTM model performed well by 39.19.7% in RMSE and 15.36% in test loss. From the statistical analysis, BiLSTM is the best performer compared with BiGRU, with a negative t-statistic (-29.65) and a very low p-value (0.00000771), indicating a statistically significant difference.

Introduction

Deep learning models are especially well-suited for temperature prediction tasks because of their capacity to learn complex representations and dependencies within data, which has led to their emergence as effective tools. These models can handle high-dimensional data, capture intricate patterns, and make predictions that traditional methods might miss, giving more knowledge of temperature trends. Despite the potential of deep learning models, selecting the appropriate architecture and configuring the model parameters remain critical challenges. Different models vary in how they preprocess data, structure their layers, and optimize learning, which can significantly impact their predictive performance. Time-series forecasting tasks like temperature prediction are best suited for recurrent neural network-based models like LSTM and GRU, built to learn temporal dependencies in sequential data.

Despite the progress in temperature prediction using deep learning models, significant gaps remain in the current literature. A major issue is the limited focus on explainability and interpretability of these models, often criticized for being "black boxes" without clear insights into their decision-making processes. Additionally, there is insufficient exploration of advanced models that combine deep learning with traditional statistical methods to enhance forecasting accuracy. Additionally, there is a need for further investigation into the generalization capabilities of DL models across various climates and geographical areas to ensure robust, globally applicable forecasting solutions.

To comprehensively evaluate the efficiency of various deep learning models for temperature prediction,

***Corresponding Author:** amritasarkar@bitmesra.ac.in

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

this study systematically compares the performance of Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM), Bidirectional GRU (BiGRU) models. Each model's setup, including data preprocessing, sequence creation, model architecture, and training configuration, is meticulously analyzed to understand its strengths and limitations.

The objectives of this research are as follows:

#Develop deep learning models for temperature prediction using environmental data

Study the impact of loss value on model access

Compare the model's performance using RMSE and loss graph

Conduct statistical analysis of models

Background

Temperature prediction is critical to weather forecasting, impacting various sectors, including agriculture, energy, and public safety, with intricate environmental data. Nowadays, deep learning (DL) has demonstrated encouraging results in improving the precision of temperature predictions. This literature review synthesises the key findings and methodologies from notable studies in this domain. Temperature prediction is among the time series forecasting tasks in which deep learning models, mainly neural networks, have shown higher performance. These models can acquire complex patterns and dependencies in data, making them appropriate for handling the non-linear and dynamic nature of weather variables.

Traditional and Machine Learning Approaches

Numerical Weather Prediction (NWP) models have been the standard for weather forecasting, but initialisation data often limits their accuracy (Dudhia, 2008; Bi et al., 2023; Waqas et al., 2024). To overcome these limitations, Gadekallu et al. (2009) utilised data mining techniques like decision trees and Random Forests, demonstrating their effectiveness in weather prediction when ample data is available (Kaur et al., 2024; Haroon et al., 2024). Similarly, Madan et al. (2018) showed that machine learning models, for example, linear regression and support vector machines, could enhance forecast accuracy over multiple days when combined with Big Data. Simple machine learning models like Linear regression techniques are inadequate for predicting time-dependent air pollution data. In contrast, Seasonal Autoregressive Integrated Moving Average (SARIMA) and Prophet models are more effective for forecasting future levels of pollutants (Samal et al., 2019). However, a more effective method for forecasting temperature is Neural networks. Basic Neural network models MLP and RBF can even realistically forecast air temperature (Francik and Kurpaska, 2020; Tran et al., 2021; Hanoon et al., 2021). As a Recurrent Neural Network, the Nonlinear Autoregressive Exogenous (NARX) NN method is superior for one-day-ahead forecasting of multiple weather attributes by capturing dynamic and chaotic weather behaviour. Biswas et al. (2014) compared to Case-Based Reasoning (CBR) methods.

Deep Learning Models

Deep learning models, especially neural networks, have proven considerable promise in temperature prediction. Long Short-Term Memory (LSTM) networks can capture long-term dependencies in time-series data due to its widespread adoption (Park et al., 2019; Karunakar et al., 2023; G et al., 2024; Waqas and Humphries, 2024). For instance, LSTM models have effectively improved forecast accuracy over short and extended periods by incorporating missing data refinement processes. Subashini et al. (2019) proposed LSTM-based neural networks trained on various weather parameters, which can effectively predict weather conditions by analyzing complex patterns in climate data. Merged models like LSTM-ARIMA have further enhanced temperature forecasting by leveraging the strengths of both models (Salman et al., 2018). Weyn et al. (2020) developed a Deep Learning Weather Prediction (DLWP) model that outperformed traditional models like the ECMWF IFS in accuracy and computational speed. Similarly, Singh et al. (2023) proposed a modified DLWP model for global precipitation prediction, which could be adapted for temperature forecasting. These models highlight the potential of deep learning in handling complex weather phenomena and rare events.

BiLSTM models outperform ordinary LSTM models in predictive performance by analysing data in a forward and backward manner (Siami-Namini et al., 2019). Alharbi and Csala (2021) showed that BiLSTM models perform better than other neural networks in short-term predictions of weather variables, such as temperature, due to their enhanced feature extraction capabilities. Liang et al. (2021) used BL-FC, based on a Bidirectional Long Short-Term Memory (Bi-LSTM) Network, for temperature modelling and forecasting. The BL-FC model outperforms the LSTM, GRU and RNN models in terms of short-term temperature prediction accuracy.

Challenges and Future Directions

The distinct characteristics of meteorological data and the absence of specific benchmarks make it difficult to completely replace conventional models with deep

DOI: https://doi.org/10.52756/ijerr.2025.v47.002

learning, notwithstanding these developments (Schultz et al., 2021). Future research should incorporate physical knowledge into models, utilise observational data, improve rare event prediction and develop hybrid models combining NWP with deep learning techniques (Bochenek and Ustrnul, 2022).

Several studies have compared the performance of DL models against traditional methods and other machine learning techniques. Siami-Namini et al. (2019) and Schultz et al. (2021) evaluated LSTM, BiLSTM and other DL models, concluding that DL models outperform traditional approaches in accuracy and robustness.

The literature review covers a range of studies focused on weather and environmental forecasting using various deep learning (DL) models, such as LSTM, BiLSTM. These studies show how well DL models work with temporal dependencies, enhancing the precision of weather forecasts and enhancing missing data management. These also highlight the superiority of DL approaches over traditional numerical weather prediction systems in capturing complex spatial-temporal patterns and non-linear relationships in the data. However, the studies also note challenges, such as the need for high-quality data, model interpretability issues, and ensuring generalization across different geographic regions and climate conditions.

Methodology

This section explicitly explains the proposed model for temperature prediction. This study assessed the performance of different deep-learning models for temperature prediction using environmental data. Error analysis for each model is examined, and the proposed model is refined.

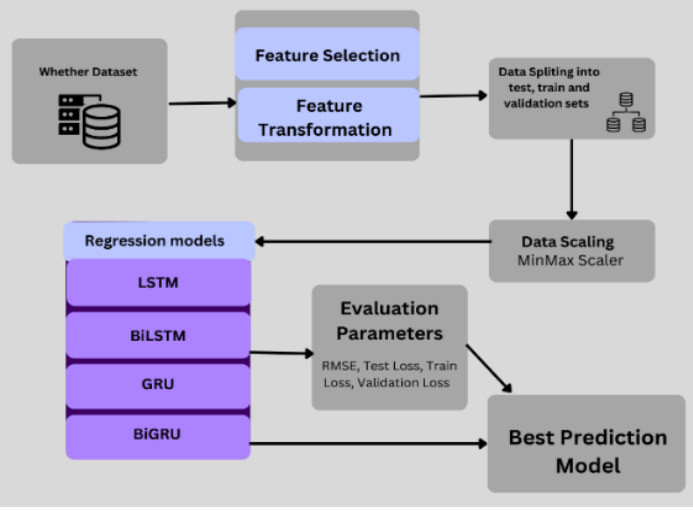


Figure 1. Proposed deep learning approach for temperature prediction.

Data Collection and Preprocessing

In the present research, data from trusted websites (Visual Crossing, 2024) was collected from 1st January 2014 to 31st July 2024 to conduct studies on temperature prediction. The collection includes 3864 meteorological observations from Ranchi, India, with 16 variables relevant to temperature prediction. These variables include dew point, humidity, precipitation, precipitation cover, wind gust, wind speed, wind direction, sea level pressure, cloud cover, visibility, solar radiation, solar energy, UV index, moon phase, and weather conditions. Temperature is the dependent numerical variable used for temperature prediction and is the average temperature in an area. Every variable is numerical in type except weather conditions, which is divided into four categories where 0 represents Clear, 1 represents Partially Cloudy, 2 represents Rain with Overcast, and 3 represents Rain with Partially Cloudy. The meteorological dataset is free of outliers and missing values. To provide a homogenous range of qualities, data is therefore normalised using the Z score and min–max scaler.

Proposed models

The methodology for temperature prediction in this study involved applying and comparing multiple deep-learning models. This section explicitly explains the proposed model for temperature prediction. This study evaluated the performance of various deep-learning models for temperature prediction using environmental data. Error analysis for each model is examined, and statistical analysis is performed to validate the best-proposed model. We employed Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM), Bidirectional GRU (BiGRU**).** For each model, they trained it on the available temperature data and used it to make predictions. The performance of each model was then evaluated using Root Mean Square Error (RMSE) as the primary metric. Additionally, scatter plots were generated for each model, comparing predicted temperatures against actual temperatures. These plots visually represented each model's accuracy and any potential biases or patterns in the predictions. We have also plotted loss function graphs for each deep learning model to predict temperature so that it quantifies the difference between the model's predicted values and the actual temperature values. The analysis of loss function is essential for training, optimizing, evaluating, and adapting deep learning models to effectively predict temperature, ensuring that they learn to make accurate and reliable predictions. We analyzed the metrics, scatter plots and loss function

DOI: https://doi.org/10.52756/ijerr.2025.v47.002

graphs to assess and compare the effectiveness of each model in predicting temperatures accurately. We also performed statistical significance testing for the models' performance metrics, so that we can analyze if the differences in RMSE among the models are statistically significant. For this, we have applied a paired t-test method. This analysis helps us understand the consistency and reliability of each model's performance.

Deep Learning Model

LSTM networks are a kind of RNN that may determine long-term dependencies by controlling the input flow using gating methods. To capture context from both past and future states, BiLSTM enhances LSTM by processing data in both forward and backward orientations. The key components of LSTM and BiLSTM models for this research are shown in Figure 2 and figure 3 respectively.

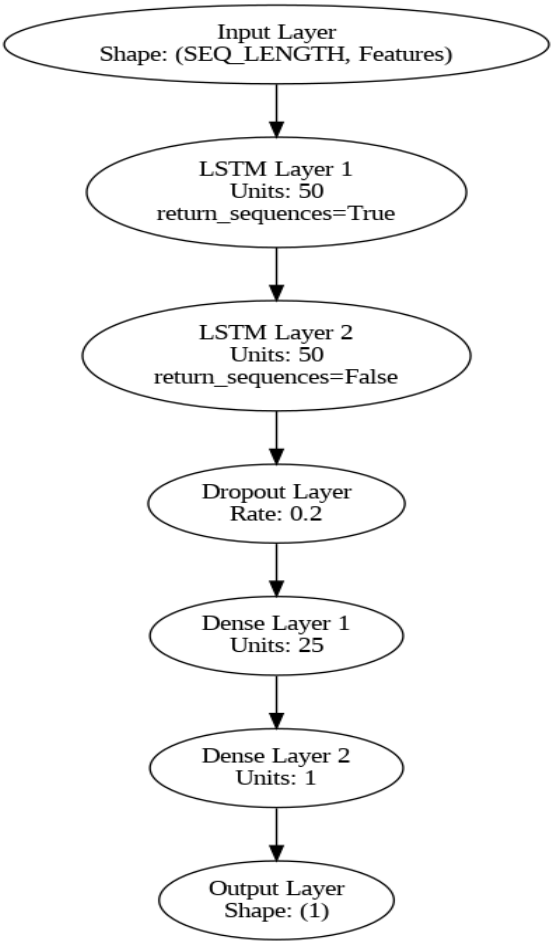


Figure 2. Proposed architecture of LSTM

GRU combines the input and forgets gates into one update gate, making it a less complicated and gate-intensive alternative to LSTM. By processing the sequence in both forward and backward directions, BiGRU expands on GRU. The fey factors of GRU and BiLGRU models for this experimental analysis shown in Figure 4 and Figure 5.

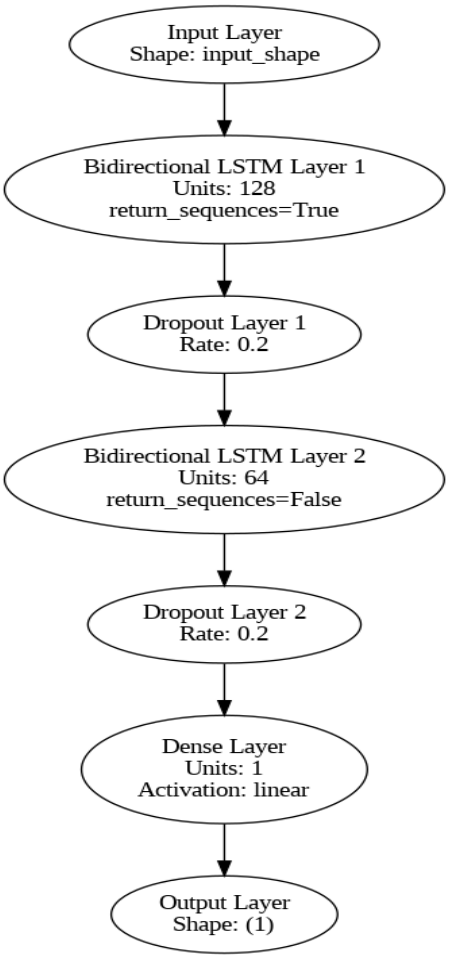


Figure 3. Proposed architecture of BiLSTM.

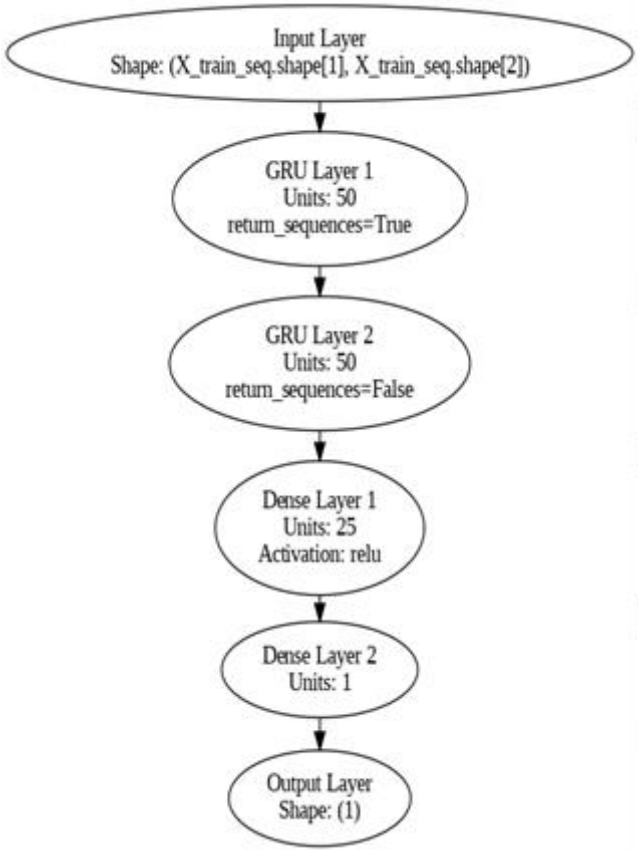


Figure 4. Proposed architecture of GRU.

DOI: https://doi.org/10.52756/ijerr.2025.v47.002

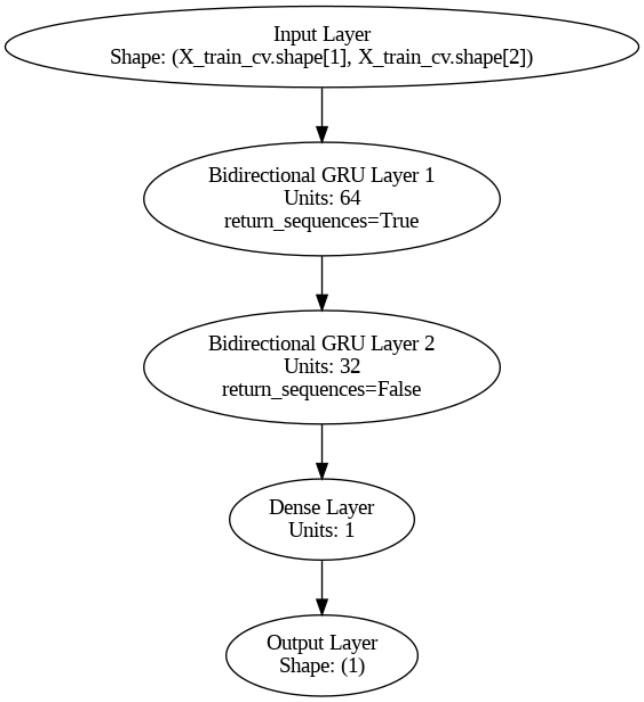


Figure 5. Proposed architecture of BiGRU.

Results and evaluation

In the search for accurate temperature prediction, various deep-learning models were evaluated to determine their effectiveness and reliability. This evaluation involves analyzing how each model processes data, structures its architecture, and undergoes training.

Experimental Setup

The experiments were run on Google Colab Environment. Table 1 provides a detailed comparison of different deep learning models (LSTM, BiLSTM, GRU and BiGRU) for temperature prediction, focusing on their Data Preprocessing, Sequence Creation, Data Reshaping, Train-Test Split, Model Architecture and Model Compilation and Training. This table also helps to understand the nuances and differences in their setup and performance.

Performance analysis of models based on visual representation of scatter plot

An “actual vs predicted” graph is a visual representation used in regression analysis to compare the values predicted by a model to the actual observed values. It helps in evaluating how well the model has predicted the target variable. It provides a visual indication of the model's performance, showing where predictions match actual values and where they deviate. We have made scatter plots in Fig. 6 of various deep-learning models (LSTM, BiLSTM, GRU and BiGRU) for predicting temperature.

Below is Table 2 where we have drawn the inference based on visual representation of the actual vs predicted graph. This table summarizes the performance of various deep learning models (LSTM, BiLSTM, GRU and BiGRU) in predicting temperatures.

The scatter plot analysis of various deep learning models (LSTM, BiLSTM, GRU and BiGRU) for temperature prediction shows that all models generally make accurate predictions, with most data points closely clustered around the diagonal line, indicating high prediction accuracy. The LSTM and GRU models exhibit low error margins, although the GRU has a few outliers that suggest potential prediction limitations. The BiLSTM model performs exceptionally well in the mid-temperature range (15°C to 30°C) but shows some deviations at higher temperatures, leading to occasional prediction errors. The BiGRU model demonstrates consistent accuracy across all temperature ranges with dense clustering around the diagonal line and minimal outliers, indicating strong predictive performance.

The LSTM model maintains high accuracy in predicting temperature patterns but slightly increases prediction error as temperatures rise, indicating slightly reduced accuracy at higher temperature ranges. The BiLSTM model similarly struggles with higher temperatures, sometimes leading to under or over-predictions. The GRU model generally provides accurate predictions with low variance but encounters occasional difficulties with extreme temperature values, reflected in a few outliers. In contrast, the BiGRU model effectively handles both central and extreme temperatures with minimal discrepancies, showcasing its strong performance across the entire temperature range.

Analysis of loss graph for Temperature Prediction

Analysing a loss graph for temperature prediction generally involves examining how well a deep-learning model learns to predict temperature values over time. The loss graph plots training loss and validation loss against the number of epochs, providing insights into the model's performance and learning behaviour. We have plotted the loss graph in Figure 7 of various deep learning models (LSTM, BiLSTM, GRU and BiGRU) for predicting temperature.

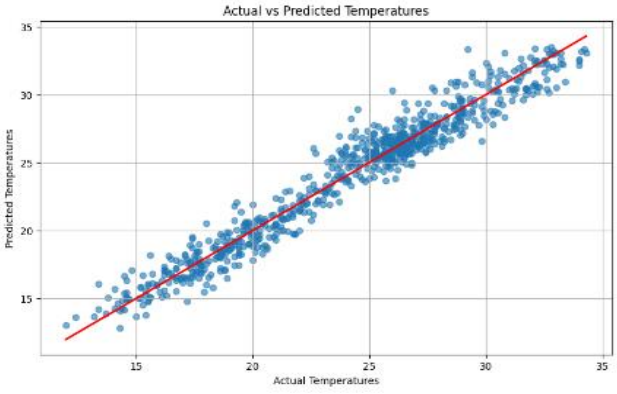
We have analysed all the loss graphs over multiple epochs and prepared Table 3 to describe the performance of various deep learning models (LSTM, BiLSTM, GRU, BiGRU) in temperature prediction. Each model demonstrates effective learning during training, with steady decreases in training and validation loss. The inferences highlight that while all models show good generalization during the early epochs, overfitting occurs at different points, particularly as the models stabilize.

DOI: https://doi.org/10.52756/ijerr.2025.v47.002

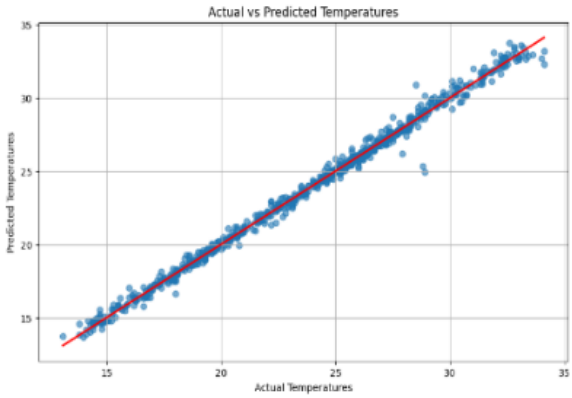
Table 1. Comprehensive overview of deep learning models applied for temperature prediction task.

Model	Data Preprocessing/ Normalization	Sequence Creation / Data Reshaping	Train-Test Split	Model Architecture	Model Compilation and Training
LSTM	Features and target normalized using MinMaxScaler.	**Sequence Creation:** Sequences of length 50 were created for training.	80-20 split between training and testing sets.	• **LSTM Layers:** Two LSTM layers; the first returns sequences, the second outputs only the last hidden state. • Dropout: 20% dropout layer added. • Dense Layers: Two Dense layers with final layer predicting temperature.	• **Loss Function**: Compiled with mean_squared_error. • **Optimizer**: Adam optimizer. • **Training**: 100 epochs with a batch size of 32. 20% of training data used as validation set.
BiLSTM	Features normalized using MinMaxScaler.	**Reshaping**: Input reshaped to match LSTM requirements, expanding data to include a time step dimension.	80-20 split; training data further split to use 16% for validation .	• **Bidirectional LSTM Layers**: Two layers, the first with 64 units and returns sequences, the second with 32 units and outputs the final hidden state. • **Dense Layer**: Final Dense layer with one neuron to predict temperature.	• **Loss Function**: Compiled with mean_squared_error. • **Optimizer**: Adam optimizer. • **Training**: 100 epochs with a batch size of 32. 20% of training data used as validation set.
GRU	Features scaled using MinMaxScaler.	**Sequence Creation**: Sequences of length 10 created for training.	80-20 split; training data further split to use 16% for validation	• **GRU Layers**: Two GRU layers, both with 50 units; the first returns sequences, the second returns the final hidden state. • **Dense Layers**: One Dense layer with 25 units (ReLU activation) and final layer with one neuron to predict temperature.	• **Loss Function**: Compiled with mean_squared_error. • **Optimizer**: Adam optimizer. • **Training**: 50 epochs with a batch size of 32. Validation set used to monitor performance after each epoch.
BiGRU	Features scaled using MinMaxScaler.	**Reshaping**: Data reshaped into a 3D array for GRU input with time steps set to 1.	80-20 split; training data further split to use 20% for validation	• **Bidirectional GRU Layers**: Two layers, the first with 64 units (returns sequences), the second with 32 units (outputs the final hidden state). • **Dense Layer**: Final Dense layer	• Loss Function: Compiled with mean_squared_error. • Optimizer: Adam optimizer. • Training: 50 epochs with a batch size of 32. Validation set used to monitor performance after each epoch.

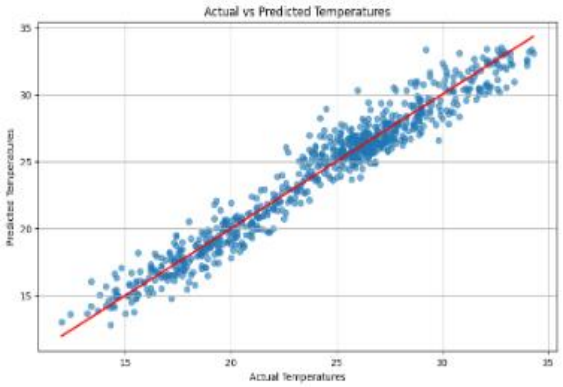
DOI: https://doi.org/10.52756/ijerr.2025.v47.002



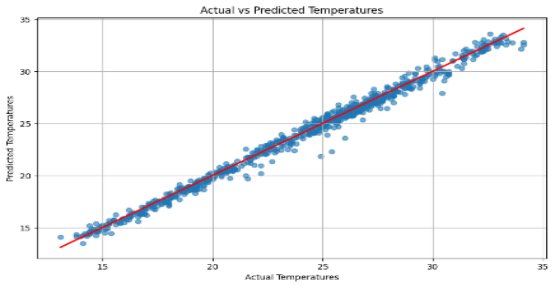
a) LSTM Model



b) BiLSTM



c) GRU



d) BiGRU

Figure 6. Visual representation of scatter plots of deep learning models.

The table emphasizes the models' strong learning capabilities and ability to capture underlying data patterns effectively.

Key observations highlight the training and validation loss trends for various deep-learning models used in temperature prediction in Table 3. All models demonstrate effective learning in the initial training phase, marked by a sharp decline in training and validation loss. As training progresses, the models reach a plateau in validation loss, indicating the onset of overfitting. The LSTM and CNN models begin overfitting after approximately 20 epochs, whereas the BiLSTM, GRU, and BiGRU models show signs of overfitting earlier, around 10 epochs. Despite this, the consistent decrease in training loss across all models shows their ability to learn effectively from the data, with the models maintaining relatively close alignment between training and validation loss before overfitting occurs. This suggests that these models are generally well-suited for temperature prediction but may require early stopping or regularization strategies to mitigate overfitting and improve generalization.

Analysis of performance considering both RMSE and loss graph

We have compared the performance of various deep learning models for temperature prediction, focusing on RMSE, test loss, training loss, validation loss, and overall inference in Table 4.

The data analysis in table 4 reveals that the BiLSTM model emerges as the top performer with the lowest RMSE, reflecting high accuracy and balanced loss values across the training, validation, and test sets, indicating strong generalization capabilities and minimal overfitting. The BiGRU model follows closely, providing good accuracy. Conversely, the LSTM model, despite maintaining low training and test losses, registers the highest RMSE, indicating poor predictive accuracy. Lastly, the GRU model underperforms with the highest RMSE and severe overfitting, leading to poor generalization. BiGRU and BiLSTM models appear most promising for temperature prediction in this context due to their superior RMSE performance and better generalisation than LSTM and GRU models.

Ranking of Deep Learning models

We have prepared a table showing the ranking of different models (BiGRU, BiLSTM, LSTM and GRU) based on their performance in temperature prediction. Table 5 compares the models using four key metrics: RMSE, test loss, training loss, and validation loss. The table then ranks each model in these categories, providing

DOI: https://doi.org/10.52756/ijerr.2025.v47.002

Table 2. Summarization of the performance of various deep learning models.

Model	Performance Analysis	Inference
LSTM	- Points are strongly clustered around the diagonal line. - High prediction accuracy with a low error margin. - Effectively captures the underlying temperature patterns.	- Highly accurate in capturing temperature patterns. - Performance slightly decreases at higher temperature ranges.
BiLSTM	- Strong predictive accuracy within the mid-temperature range (15°C to 30°C). - Tight clustering around the diagonal line in this range. - Some deviation at higher temperatures, indicating occasional prediction errors.	- Performs well in mid-temperature ranges. - Shows some prediction errors at higher temperatures.
GRU	- Tight clustering of points around the diagonal line. - Indicates accurate predictions with low variance. - A few outliers suggest potential limitations.	- Generally accurate with low prediction variance. - Occasional challenges in predicting extreme temperatures.
BiGRU	- Dense clustering around the diagonal line. - Very accurate predictions with minimal variance. - Consistently performs well across a broad temperature range.	- Consistently accurate across different temperature ranges. - Effectively handles both central and extreme temperatures.

Table 3. Summarize the analysis of the loss graph for different deep-learning models.

Model	No of Epochs	Train Loss	Validation Loss	Inference
LSTM	Loss graph over 100 epochs	Decreases steadily throughout the training process, indicating continuous learning.	Decreases sharply initially, then stabilizes after around 20 epochs.	The model effectively captures the underlying patterns, with minimal overfitting observed only after 20 epochs.
BiLSTM	Loss graph over 100 epochs	Decreases steadily throughout the training process, showing effective learning.	Decreases sharply initially, then stabilizes after around 10 epochs.	The model quickly learns the data patterns, maintaining good generalization up to 10 epochs.
GRU	Loss graph over 50 epochs	Decreases steadily throughout training, indicating effective learning.	Decreases sharply initially, then stabilizes after around 10 epochs.	The model shows strong learning capabilities, with good performance before slight overfitting begins.
BiGRU	Loss graph over 50 epochs	Decreases steadily throughout training, showing effective learning.	Decreases sharply initially, then stabilizes after around 10 epochs.	The model learns efficiently, with excellent generalization during the initial epochs.

DOI: https://doi.org/10.52756/ijerr.2025.v47.002

Table 4. Comparative table on RMSE, Test Loss, Training Loss, and Validation Loss function.

Model	RMSE	Test Loss	Train Loss	Validation Loss	Inference
BiLSTM	0.3919	0.1536	0.1154	0.1451	The best performance with the lowest RMSE and low loss values indicates excellent predictive accuracy and minimal overfitting.
BiGRU	0.4870	0.2372	0.2420	0.2629	Good performance with slightly higher RMSE and balanced loss values, showing strong generalization capabilities.
LSTM	1.1290	0.2022	0.0019	0.0022	Very low loss values but highest RMSE, indicating overfitting and poor predictive accuracy on unseen data.
GRU	1.1137	1.2404	1.1859	1.3460	Poor performance with high RMSE and noticeable overfitting leads to poor generalization.

Table 5. Ranking of deep learning models by each metric.

Model	RMSE Ranking	Test Loss Ranking	Training Loss Ranking	Validation Loss Ranking	Overall Ranking
BiLSTM	1	1	2	2	1
BiGRU	2	2	3	3	2
LSTM	4	3	1	1	3
GRU	3	4	4	4	4

Table 6. RMSE values for deep learning models across five cross-validation folds.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Observation	Inference
LSTM	1.1825	1.1477	1.1277	1.1905	1.1460	The RMSE values range from 1.1277 to 1.1905 across the five folds.	Indicates relatively consistent performance with a moderate error level.
BiLSTM	0.0706	0.0739	0.0647	0.0761	0.0675	The RMSE values are very low, ranging from 0.0647 to 0.0761.	Suggests that the BiLSTM model has the lowest prediction errors and performs best among all models.
GRU	1.1731	1.3266	1.2332	1.1712	1.2275	The RMSE values vary between 1.1712 and 1.3266.	Shows more variability compared to LSTM but with a similar range of moderate errors.
BiGRU	0.6253	0.5450	0.6062	0.5682	0.5448	The RMSE values range from 0.5448 to 0.6253.	Higher errors than BiLSTM but significantly lower than LSTM and GRU, indicating better performance than unidirectional models but worse than BiLSTM.

DOI: https://doi.org/10.52756/ijerr.2025.v47.002

Table 7. Performance Comparison of Deep Learning Models Using t-Statistic and p-Value.

Pair	t-statistic	p-value	Observation	Inference
LSTM X BiLSTM	104.13358312	0.00000005	The t-statistic is high (104.13), and the p-value is extremely low (0.00000005).	This suggests that the BiLSTM model significantly outperforms the LSTM model.
LSTM X GRU	-1.81879457	0.14307733	The t-statistic is -1.82, and the p-value is 0.1431, higher than the typical 0.05 threshold.	This suggests no statistically significant difference between the LSTM and GRU models.
LSTM X BiGRU	31.75092197	0.00000586	The t-statistic is 31.75, and the p-value is very low (0.00000586).	This indicates that the BiGRU model significantly outperforms the LSTM model.
BiLSTM X GRU	-40.74289141	0.00000217	The t-statistic is -40.74 with a very low p-value (2.17E-06).	The negative value suggests that the GRU performs worse than the BiLSTM.
BiLSTM X BiGRU	-29.64872852	0.00000771	The t-statistic is -29.65, and the p-value is very low (7.71E-06).	This indicates a significant difference, with the BiGRU performing worse than the BiLSTM.
GRU X BiGRU	16.32689207	0.00008237	The t-statistic is 16.33 with a very low p-value (8.24E-05).	This suggests that the BiGRU model outperforms the GRU model.

Table 8. Ranking of deep learning models by statistical analysis.

Model	Rank	Performance Summary
BiLSTM	1	Outperforms LSTM, GRU, and BiGRU models significantly.
BiGRU	2	It Outperforms LSTM and GRU but performs worse than BiLSTM.
LSTM	3	LSTM and GRU have no significant difference; they perform worse than BiLSTM and BiGRU.
GRU	4	Performs worse than BiLSTM and BiGRU; no significant difference from LSTM.

an overall ranking based on their performance across these metrics. The Overall Ranking is derived by summing the individual rankings for each model across the main two metrics (RMSE, test loss).

The ranking Table 5 reveals that the BiLSTM model is consistently a top performer among the others, achieving the best overall ranking due to its lowest RMSE, test loss, and favourable training and validation loss values. The BiLGRU model follows closely, with solid performance across all metrics. The LSTM model, while having the lowest training and validation losses, ranks third overall due to a higher RMSE, indicating less accurate predictions. The GRU model shows moderate performance but is hindered by higher losses, placing it fourth.

Statistical analysis of Deep Learning models

We have further employed statistical comparisons between different deep-learning models (LSTM, BiLSTM, GRU, and BiGRU) for temperature prediction based on environmental data. Before doing that, we applied all deep learning models across five folds of cross-validation and computed Root Mean Square Error (RMSE) values for each fold. Table 6 shows the RMSE values for five deep-learning models across five cross-validation folds. RMSE measures the average magnitude of errors in a set of predictions, with lower values indicating better model performance. These RMSE values provide a basis for further statistical analysis, such as conducting t-tests to compare the performance of these models rigorously. Comparing RMSE values of models across the folds using t-tests can determine if the observed differences in model performance are statistically significant.

According to the RMSE value, Table 6 shows low RMSE values for BiLSTM, we expect t-tests comparing BiLSTM with the other models to show statistically

significant differences, particularly against CNN and the unidirectional models (LSTM and GRU). BiGRU is also likely to improve over GRU and CNN significantly, but less so than BiLSTM. The table highlights that BiLSTM is the most accurate model for temperature prediction with consistently low RMSE values. The RMSE values set the stage for statistical testing to confirm the significance of these differences.

We have also prepared Table 7, which shows the statistical comparisons between different deep-learning models used for temperature prediction based on environmental data. To determine whether there are statistically significant differences in the models' performance, these comparisons are conducted using the t-statistic and p-value. According to the variation in sample data, the t-statistic calculates the size of the difference. A higher absolute t-value indicates a greater discrepancy between the models. The p-value indicates the probability that the observed difference between models occurred by chance. In most cases, a p-value of less than 0.05 indicates the statistical significance of the model difference.

Based on the t-statistics and p-values provided in Table 7, we can conclude that there are significant differences in performance among the LSTM, BiLSTM, GRU, and BiGRU models. The BiLSTM model significantly outperforms the LSTM and GRU models, as indicated by the high t-statistics (104.13 and -40.74, respectively) and extremely low p-values. Similarly, the BiGRU model significantly outperforms both the LSTM and GRU models, with very low p-values for both comparisons. However, there is no statistically significant difference between the LSTM and GRU models, as evidenced by a p-value higher than the typical 0.05 threshold.

Overall, according to the analysis from Table 8, models incorporating bidirectional architectures (BiLSTM and BiGRU) consistently demonstrate superior predictive performance compared to their unidirectional counterparts. From this analysis, the negative t-statistic between BiLSTM and BiGRU implies that the BiGRU model performs worse than the BiLSTM model. Based on statistical analysis, BiLSTM emerges as the top performer for temperature prediction tasks for environmental data of Ranchi, India.

Conclusion

This work explored temperature prediction using advanced deep learning models, with particular attention to how well different architectures work. Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Gated Recurrent Unit (GRU), and Bidirectional GRU (BiGRU) handle the intricate, nonlinear relationships present in climatic data. The research attempted to ascertain which deep learning technique is most appropriate for precise temperature forecasting by contrasting different models. Performance indicators like Root Mean Square Error (RMSE), visual scatter plots and loss graphs were used in the evaluation to review each model's predictive power thoroughly.

The findings reveal that different deep learning models excel in different aspects of temperature prediction. Temperature patterns over long periods of time can be predicted using LSTM and BiLSTM models because they are particularly good at capturing long-term dependencies in sequential data. GRU and BiGRU models, with their simpler architecture, offer competitive accuracy with reduced computational complexity, making them a practical choice for real-time forecasting applications. The research highlights that no single model outperforms others in all scenarios. The nature of the data, the intended prediction horizon, and the available computing power should all be considered when selecting a model for a given forecasting task.

According to the analysis, BiLSTM performs best for temperature prediction based on statistical significance testing, RMSE and loss function analysis. BiLSTM shows better average predictive accuracy and generalization and more statistically consistent performance across multiple evaluations. BiLSTM better serves most real-world applications that aim to reduce prediction error and achieve balanced performance across environmental datasets for temperature prediction. However, BiLSTM could be the suitable option if statistical consistency and low variance in predictions are more critical for the task.

This research aims to improve our knowledge of the many deep-learning architectures that may be applied to temperature prediction. This knowledge will be useful for future investigations and applications in climate science and other areas where precise forecasting is essential. In addition to highlighting the potential of deep learning models to advance temperature prediction, this thorough investigation also underscores the necessity of careful model selection and configuration for optimal outcomes.

Future research should address these gaps by developing methods to enhance model interpretability and transparency, integrating physical and meteorological knowledge into deep learning frameworks, and creating more versatile models that generalize across diverse geographic and climatic conditions. Furthermore, adding

DOI: https://doi.org/10.52756/ijerr.2025.v47.002

domain knowledge and focusing more on localized and extreme weather event prediction could boost the precision and usefulness of deep learning models for temperature forecasting. Overcoming these obstacles will enable the field to get closer to utilizing deep learning's full potential for improving weather prediction, which will ultimately lead to more informed decisions being made in industries like public safety, energy, and agriculture.

Acknowledgement

We would like to express our sincere gratitude to Birla institute of Technology, Mesra (Lalpur Unit) for providing the resources and support necessary to conduct this research.

Conflict of interest

The authors declare no conflict of interest

References

Alharbi, F. R., & Csala, D. (2021). Wind speed and solar irradiance prediction using a bidirectional long short-term memory model based on neural networks. *Energies*, *14*(20), 6501. https://doi.org/10.3390/en14206501

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3

Biswas, S. K., Sinha, N., Purkayastha, B., & Marbaniang, L. (2014). Weather prediction by recurrent neural network dynamics. *International Journal of Intelligent Engineering Informatics*, *2*(2/3), 166–180. https://doi.org/10.1504/IJIEI.2014.066208

Bochenek, B., & Ustrnul, Z. (2022). Machine learning in weather prediction and climate analyses—Applications and perspectives. *Atmosphere*, *13*(2), 180. https://doi.org/10.3390/atmos13020180

Dudhia, J. (2008). Fundamental issues in numerical weather prediction. *Large-Scale Disasters*, pp. 447–452. https://doi.org/10.1017/cbo9780511535963.021

Francik, S., & Kurpaska, S. (2020). The Use of Artificial Neural Networks for Forecasting of Air Temperature inside a Heated Foil Tunnel. *Sensors*, *20*, 652. https://doi.org/10.3390/s20030652

G, T., MHM, K. P., & Rao P, S. (2024). An Improved Parallel Heterogeneous Long Short-Term Memory Model with Bayesian Optimization for Time Series Prediction. *International Journal of Experimental Research and Review*, *45*(Spl Vol), 106–118. https://doi.org/10.52756/ijerr.2024.v45spl.009

Gadekallu, T. R., Kidwai, B., Sharma, S., Pareek, R., & Karnam, S. (2019). Application of data mining techniques in weather forecasting. In *Sentiment Analysis and Knowledge Discovery in Contemporary Business*, pp. 162–174. IGI Global. https://doi.org/10.4018/978-1-5225-4999-4.ch010

Hanoon, M.S., Ahmed, A.N., Zaini, N., Razzaq, A., Kumar, P., Sherif, M., Sefelnasr A., & El-Shafie, A. (2021). Developing machine learning algorithms for meteorological temperature and humidity forecasting at Terengganu state in Malaysia. *Sci Rep.*, *11*(1), 18935. https://doi.org/10.1038/s41598-021-96872-w

Haroon, M., Siddiqui, Z. A., Husain, M., Ali, A., & Ahmad, T. (2024). A Proactive Approach to Fault Tolerance Using Predictive Machine Learning Models in Distributed Systems. *International Journal of Experimental Research and Review*, *44*, 208–220. https://doi.org/10.52756/*IJERR*.2024.v44spl.018

Karunakar, G., Shreya, G., Shanmukha Priya, G., Vasavi, A., Sriharshini, G., & Mahalakshmi, G. (2023). Unified time series analysis with Bi-long short-term memory model for early prediction of dyslipidemia in steel workers. *E-Prime - Advances in Electrical Engineering, Electronics and Energy*, *6*, 100302. https://doi.org/10.1016/j.prime.2023.100302

Kaur, S., Bansal, K., & Kumar, Y. (2024). Machine Learning-Based Prediction System for Risk Assessment of Hypertension Using Symptoms Investigations. *International Journal of Experimental Research and Review*, *46*, 139–149. https://doi.org/10.52756/ijerr.2024.v46.011

Liang, S., Wang, D., Wu, J., Wang, R., & Wang, R. (2021). Method of bidirectional LSTM modelling for the atmospheric temperature. *Intelligent Automation & Soft Computing*, *29*(3), 701–714. https://doi.org/10.32604/iasc.2021.020010

Madan, S., Kumar, P., Rawat, S., & Choudhury, T. (2018). Analysis of Weather Prediction using Machine Learning & Big Data. *2018 International Conference on Advances in Computing and Communication Engineering* (ICACCE), pp. 259–264. https://doi.org/10.1109/icacce.2018.8441679

Park, I., Kim, H. S., Lee, J., Kim, J. H., Song, C. H., & Kim, H. K. (2019). Temperature prediction using the missing data refinement model based on a long short-term memory neural network. *Atmosphere, 10*(11), 718. https://doi.org/10.3390/atmos10110718

Salman, A. G., Heryadi, Y., Abdurahman, E., & Suparta, W. (2018). Weather forecasting using merged long short-term memory model (LSTM) and autoregressive integrated moving average (ARIMA) model. *Journal of Computer Science, 14*(7), 930–938. https://doi.org/10.3844/jcssp.2018.930.938

Samal, K. K. R., Babu, K. S., Das, S. K., & Acharya, A. (2019). Time series-based air pollution forecasting using SARIMA and Prophet model. In *Proceedings of the 2019 International Conference on Information Technology and Computer Communications*, pp. 88–92. https://doi.org/10.1145/3355402.3355417

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., & Stadtler, S. (2021). Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379*(2194), 20200097. https://doi.org/10.1098/rsta.2020.0097

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). The performance of LSTM and BiLSTM in forecasting time series. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 3285–3292. https://doi.org/10.1109/BigData47090.2019.9005997

Singh, M., Acharya, N., Patel, P., Jamshidi, S., Yang, Z.-L., Kumar, B., Rao, S., Gill, S. S., Chattopadhyay, R., Nanjundiah, R. S., & Niyogi, D. (2023). A modified deep learning weather prediction using cubed sphere for global precipitation. *Frontiers in Climate, 4*. https://doi.org/10.3389/fclim.2022.1022624

Subashini, A., Thamarai, S. M., & Meyyappan, T. (2019). Advanced weather forecasting prediction using deep learning. *International Journal of Research in Applied Science and Engineering Technology, 7*(8), 939–945. https://doi.org/10.22214/ijraset.2019.8139

Tran, T. T. K., Bateni, S. M., Ki, S. J., & Vosoughifar, H. (2021). A review of neural networks for air temperature forecasting. *Water, 13*(9), 1294. https://doi.org/10.3390/w13091294

Waqas, M., & Humphries, U. W. (2024). A critical review of RNN and LSTM variants in hydrological time series predictions. *MethodsX, 13*, 102946. https://doi.org/10.1016/j.mex.2024.102946

Visual Crossing. (2024). Weather data services. Retrieved June 20, 2024, from https://www.visualcrossing.com/weather/weather-data-services

Waqas, M., Humphries, U. W., Chueasa, B., & Wangwongchai, A. (2024). Artificial intelligence and numerical weather prediction models: A technical survey. *Natural Hazards Research*. https://doi.org/10.1016/j.nhres.2024.11.004

Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems, 12*(9), e2020MS002109. https://doi.org/10.1029/2020ms002109

How to cite this Article:

Amrita Sarkar, Vandana Bhattacharjee and Prachi Pandey (2025). An Empirical Analysis of Deep Learning Models for Temperature Prediction. *International Journal of Experimental Research and Review*, *47*, 12-24.

DOI : https://doi.org/10.52756/ ijerr.2025.v47.002

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.